JOURNAL OF APPLIED MEASUREMENT, *16*(3), 251-267

Using the Partial Credit Model to Evaluate the Student Engagement in Mathematics Scale

Micela Leis

Karen M. Schmidt

Sara E. Rimm-Kaufman
University of Virginia

The Student Engagement in Mathematics Scale (SEMS) is a self-report measure that was created to assess three dimensions of student engagement (social, emotional, and cognitive) in mathematics based on a single day of class. In the current study, the SEMS was administered to a sample of 360 fifth graders from a large Mid-Atlantic district. The Rasch partial credit model (PCM) was used to analyze the psychometric properties of each sub-dimension of the SEMS. Misfitting items were removed from the final analysis. In general, items represented a range of engagement levels. Results show that the SEMS is an effective measure for researchers and practitioners to assess upper elementary school students' perception of their engagement in math. The paper concludes with several recommendations for researchers considering using the SEMS.

Requests for reprints should be sent to Micela Leis, University of Virginia, P. O. Box 800784, Charlottesville, VA 22904, USA, e-mail: ml3kc@virginia.edu.

Math achievement of students in the United States is currently of great concern. The Program for International Student Assessment (PISA) reported that the United States ranked 36th among developing nations in math, a ranking that has decreased over the past decade (PISA, 2013). Efforts designed to enhance student engagement in mathematics classrooms hold promise as one way to reverse this downward trend and improve math outcomes. Engagement refers to a person's active involvement in a task or activity (Reeve, Jang, Carrell, Jeon, and Barch, 2004). Researchers have accrued evidence that student engagement is an important predictor of learning and academic performance (Christenson, Reschly, and Wylie, 2012; Fredricks, Blumenfeld, and Paris, 2004). However, a psychometrically sound measure of student engagement in mathematics is needed. There are few instruments that specifically measure student engagement in mathematics classrooms in the United States (see Kong, Wong, and Lam, 2003 for a math engagement measure for students in China).

In 2009, Rimm-Kaufman and colleagues created the Student Engagement in Mathematics Scale (SEMS) (Appendix) to measure fifth graders' engagement, as reported by students, within the broader context of the Responsive Classroom Efficacy Study (RCES), a three-year longitudinal randomized controlled trial of a social and emotional learning intervention, the *Responsive Classroom*® (RC) approach (Rimm-Kaufman, Baroody, Larsen, Curby, and Abry, 2014). The SEMS is comprised of three subscales of student-report measures of cognitive, emotional, and social engagement. The social engagement subscale was adopted in its existing form from the student-report measure of social engagement developed and used by Patrick, Ryan, and Kaplan (2007), with the only modification involving the addition of the phrase "in math class". The student-report measures of cognitive and emotional engagement were developed based upon measures created by Meece (2009); Kong et al. (2003); Rowley, Kurtz-Costes, Meyer, and Kizzie (2009); and Skinner and Belmont (1993) to assess students' report of cognitive and emotional engagement in relation to math class on a specific day. A more detailed description of the creation and validation of the SEMS can be found in the work of Rimm-Kaufman et al. (2014).

One of the unique features of SEMS is that it measures engagement on a particular day in math class with the idea that sampling students' engagement on different days produces an average level of engagement as well as variation in their engagement over time. To our knowledge, there are no existing measures that measure students' perception of their experience in math class in this manner. Research on the psychometric properties of the SEMS has been limited to applications of classical test theory (CTT), which is founded on the *true score equation* (Raykov and Marcoulides, 2011), and confirmatory factor analysis, which identified three dimensions of engagement in mathematics: cognitive engagement, emotional engagement, and social engagement (Rimm-Kaufman et al., 2014). Rimm-Kaufman and colleagues (2014) found that students in RC schools did not have statistically significant different reported engagement scores than their peers in control schools: a surprising finding, as RC training focuses on creating more engaging classrooms (Northeast Foundation for Children, 2014).

A close analysis of the psychometric properties of the scale, using a Rasch-based model (Rasch, 1960), may help explain this unexpected finding. Rasch analysis allows for a more in-depth item-level analysis of measures, error, fit, and information from the SEMS than that obtained with the *true score* equation (Embretson and Reise, 2000). Rasch analysis is also useful in examining participants' variability in the use of the response scale (Teman, 2013).

This paper presents the first instance of applying Rasch analysis to scores from the SEMS. The SEMS uses a 4-point ordered response scale from *no, not at all true* to *yes, very true*. In conducting the analysis, it is desirable to see the formation of a continuum, with participants who have less of the trait at one extreme, and participants who have more of the trait at the other extreme (Green and Frantom, 2002). Students who answered 4 (*yes, very true*) should have more engagement in mathematics than students who answered 3 (*often true*), 2 (*a little true*), or 1 (*no, not at all true*). As the responses to the SEMS items are scored in multiple ordered categories, a Rasch-based model for polytomous data, the partial credit model (PCM; Masters, 1982), was used. The PCM extends Andrich's rating scale model by allowing the response choices to vary in both number and structure from item to item (Masters, 1982). The idea of separable parameters is important because it allows us to consider varied distances in calculating fit statistics for the SEMS.

Specifically, the current study uses the PCM to conduct a comprehensive item-level analysis of scores from the SEMS. Fit statistics, which demonstrate the validity of the scores on the overall measure, were calculated and used to inform decisions about removing misfitting items from the model in order to retain only the items that helped elucidate the latent engagement trait. The PCM was also used to examine the fifth graders' use of the four-option ordered response scale.

Literature Review

Engagement in Mathematics Classrooms

Engagement has been described by researchers as, "the glue, or mediator, that links important contexts—home, school, peers, and community—to students and, in turn, to outcomes of interest" (Reschly and Christensen, 2012, p. 3). Researchers and educators agree that engagement is critical for learning and a substantial body of research exists that establishes that engagement forecasts school success. Children who stay on task, attend to learning goals, and participate actively in learning experience tend to show better academic achievement in elementary school and beyond (Fredricks et al., 2004; Greenwood, Horton, and Utley, 2002; Hughes and Kwok, 2007; Ladd and Dinella, 2009; Ponitz, Rimm-Kaufman, Grimm, and Curby, 2009). Engagement can be measured via observational, teacher-report, and student-report methods (Finn and Zimmer, 2012), with each technique offering advantages and disadvantages.

The current study focuses measuring engagement through student-report as students' appraisal of their own school experience offers important information on engagement (Rimm-Kaufman et al., 2014). The SEMS is a self-report measure. Other instruments that measure student engagement focus on general engagement of the student, such as the perceived relevance of school (Appleton, Christenson, Kim, and Reschly, 2006) or institutional emphases of good practices (Carini, Kuh, and Klein, 2006). The SEMS specifically focuses on measuring student-reported engagement in relation to a specific day of math class (Rimm-Kaufman et al., 2014). This is important because, although fifth graders are experts in *appearing* to be engaged in learning, they need to actually feel engaged in order to internalize the instruction. Understanding the engagement of students, especially in regards to the different dimensions of engagement, can help a teacher understand which aspects of engagement need to be targeted.

Dimensionality of the SEMS

The engagement literature supports the view that engagement is comprised of social, emotional, and cognitive dimensions (Fredricks et al., 2004). Social engagement refers to students' day-to-day social exchanges with peers that are tethered to the instructional content (Patrick et al., 2007). Emotional engagement refers to children's feelings of connection to content, interest in learning, and enjoyment of solving problems and thinking about content (Fredricks et al, 2004). Cognitive engagement refers to the extent to which children show a willingness to exert effort to understand content, work through difficult problems, and manage and direct their attention toward the task at hand (Christenson et al., 2012). The SEMS was created to assess these three dimensions of engagement through student self report. In a confirmatory factor analysis (CFA) of the SEMS, using data from 387 fifth graders, Rimm-Kaufman et al. (2014) found that, after deleting three poorly fitting items, the three-

factor model of social, emotional, and cognitive engagement was well-fitting (RMSEA = .03, CFI = .96, TLI = .96). All factor loadings were statistically significant ($p < .01$). Though the CFA in conjunction with reliability and item analysis based on the *true score equation* gives valuable information about the SEMS, there are several advantages of using a Rasch model for analysis.

Rasch Measurement

Rasch measurement has various desirable features (Embretson and Reise, 2000). In particular, trait level and item properties are estimated separately, as they are considered independent variables. Therefore, the latent trait ability of the individual does not depend on comparing said individual to a representative sample of the population. If the data fit the model, then the Rasch model is considered to be invariant (Bond and Fox, 2007). This means that the measurement instrument does not affect what is being measured, which is a goal of psychological measurement creation.

We had three goals in performing a Rasch analysis on the SEMS data: (a) to create a scale that fits the Rasch model; (b) assessing the quality of the 4-point ordered response scale for each of the three subscales; (c) analyzing how well the scale worked in covering the various levels of engagement in mathematics. Rasch analysis provides a comprehensive item-level analysis. Items that do not adequately fit the Rasch model are easily identified through goodness-of-fit criteria. These items can then be deleted from the model to improve model fit (Bond and Fox, 2007). Rasch analysis also allows one to assess the quality of the rating scale. For instance, diagnostic procedures can be used to assess how well the rating scale actually worked, and provide information on whether different categories should be collapsed (Bond and Fox, 2007). The Rasch model gives information on the ability of the items to tap into the various level of the trait being measured (King and Bond, 1996). If items are not distributed across all ability levels, then the researcher will not be able to precisely measure persons with ability levels outside of those measured by the items.

Partial Credit Model (PCM)

The one-parameter Rasch model is appropriate only for dichotomous data as it is a single parameter logistic model, in which the parameter refers to the location on the continuum of the item on the latent trait of interest (Green, 2002). However, the SEMS consists of four ordered categories, in which the respondent designates their level of agreement with the item from *No, not at all true* at one extreme, to *Yes, very true* at the other extreme, resulting in polytomous response options. Rather than using the Andrich's Rating Scale model, we decided to use the PCM as it extends the Rating Scale model to situations in which response choices are allowed to vary in structure and number from item to item (Masters, 1982).

Applying the PCM to the SEMS means that we are interpreting the responses in terms of *steps* that have to be taken to complete the item. Therefore, a person who chooses *often true* as a response to a statement on the SEMS can be considered to have chosen *a little true* over *no, not at all true* (first step taken) and also *often true* over *a little true* (second step taken), but to have failed to choose *yes, very true* over *often true* (third step rejected) (Masters, 1982). More recently, some researchers in the Rasch field have veered away from using the term "step," due to its being misconstrued as an active, deliberate step. We use Masters' original wording to convey the use of a scale of ordered categories. The PCM is a direct model, with the probability of person n scoring x on item i is given in a single equation:

$$\pi_{xni} = \frac{\exp\sum_{j=0}^{x}(\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}(\beta_n - \delta_{ij})}, x = 0,1,\ldots m_i. \quad (1)$$

Where, for notational convenience,

$$\sum_{j=0}^{0}(\beta_n - \delta_{ij}) \equiv 0.$$

Equation 1 gives the probability of person n scoring x on the m_i-step item i (π_{xni}) as a function of the person's position β_n on the variable and the difficulties of the m_i "steps" in item i. The observation x is a count of successfully completed item steps. The numerator of the model gives the difficulties of the x *completed* steps. The delta parameters (δ_{ij}) refers to the step difficulty, or threshold, specified per item. That is, δ_{ij} represents the location where two categories intersect on the latent-trait continuum. Higher δ_{ij} values are associated with higher trait levels, relative to other categories within an item (Masters, 1982).

Estimating category measures allows the researcher to examine each item for its usefulness in targeting students with different levels of engagement in math. In examining these category measures, we can determine if there an adequate number of items that address students that have a range of engagement levels. Another value of the PCM is that it allows one to see if there are disordered category measures, meaning that a specific response category never has the highest probability of being chosen, no matter the trait level of the respondent.

Method

Participants

The RCES enrolled 24 schools in 2008; schools were randomized into intervention ($n = 13$) and control ($n = 11$) conditions. RCES data collection efforts involved studying: third grade students and teachers in 2008-2009, fourth grade students and teachers in 2009-2010, and fifth grade students and teachers in 2010-2011. In 2010, prior to student participants' entry into fifth grade, all 24 schools were invited to participate in the fifth grade math engagement data collection effort. Twenty schools (12 intervention, 8 control) enrolled, representing an 83% response rate. All 20 schools were located in a suburban area in a mid-Atlantic state. The schools were socioeconomically and linguistically diverse (33% free/reduced price lunch [FRPL], 31% English language learners [ELL]). All fifth grade teachers in the 20 schools were recruited by the research team through in-person meetings with principals and teachers. Sixty-three teachers enrolled, corresponding to a 79% response rate.

Approximately five children in the 63 fifth grade classrooms were selected from the larger pool of RCES student participants to take the SEMS ($n = 387$). Selection was conducted randomly for each classroom bounded by two constraints: (a) blocking by gender in order to have similar numbers of male and female participants, and (b) demographic match to the whole school (based on ethnicity, FRPL, and ELL percentages). The selected students were given the SEMS at three time points over the school year. The current study uses data from the first time the student completed the SEMS during the first half of the school year, resulting in 360 students (representing a 93% response rate). Based on school records, the final sample of student participants ($n = 360$; 52% female) were on average 10.68 years old ($SD = .40$) at the beginning of the school year. Twenty-one percent of participants qualified for FRPL (defined as $40,793 for a family of four, roughly below 180% of the federal poverty guideline).

Instrumentation

The SEMS instrument is comprised of 18 items (see Appendix), each rated on a 4-point scale. For each item, students were asked to rate their level of agreement (1 = *No, not at all true*, 2 = *A little true*, 3 = *Often true*, 4 = *Yes, very true*) with each statement about a math class they had just experienced on that day. Some example statements are: "Students in my math class helped each other learn today," "I enjoyed thinking about math today," and "I did a lot of thinking in math class today." Items 2, 4, 11, and 12 are reverse scored, so that higher scores on every statement correspond with higher levels of engagement in mathematics. A pilot study of the SEMS showed a high reliability for the full measure (Cronbach's alpha = .90) (Rimm-Kaufman et al., 2014). Confirmatory factor analysis of SEMS data using a longitudinal data set (Rimm-Kaufman et al., 2014) resulted in a three-factor solution representing sub-constructs of cognitive engagement, emotional engagement, and social engagement, with Cronbach's alphas of .78, .91, and .74, respectively. These reliability standards (especially for cognitive and social engagement),

though low by conventional standards, correspond to estimates for self-report data on elementary-aged children (Griggs, Rimm-Kaufman, Merritt, and Patton, 2013; McMahon, Parnes, Keys, and Viola, 2008). The young age of the respondents, who were mainly between 10 and 11 years old, could influence their understanding of the question or response choices, which may help explain why student report data typically have lower Cronbach's alphas than adult report data.

Analysis Method

There are two criteria that must be met in applying the PCM: unidimensionality and local independence (Embretson and Reise, 2000). Unidimensionality refers to the assumption that a single latent trait variable is responsible for the variance between item responses. Local independence and unidimensionality are related by definition as a data set is considered unidimensional when item responses are locally independent due to a single latent variable (McDonald, 1981). The creation and previous analyses of the SEMS suggest that the measure is multidimensional, and therefore must be separated into different submeasures prior to Rasch analysis. As the SEMS is a relatively new measure, we first wanted to explore the factor structure of the measure and determine if all 18 items were well fitting for this data set. Therefore, data analysis consisted of two steps: (1) an exploratory factor analysis (EFA); (2) a principal component analysis (PCA) of the residuals was performed for the total scale as well as separately for each subscale of engagement identified from the EFA. Prior to analyses, data were examined for outliers (both univariate and multivariate), and normality. Outliers were removed from the data set.

Exploratory factor analysis. An EFA was conducted to determine the factor structure of the SEMS. IBM SPSS 21 software (IBM Corp, 2012) was used to conduct the EFA. Principal axis factor (PAF) extraction was used to determine the number of latent traits. For ease of interpretation, only orthogonal rotations were considered, as these constrain factors to be uncorrelated. Therefore, varimax rotations were considered in the attempt to uncover simple structure. Empirical and theoretical evidence was examined in choosing the number of factors to retain (Cattell, 1966; Horn, 1965; Kaiser, 1958). Items with poor fit were deleted from further analysis.

Rasch analysis. After deleting items with poor fit, PCMs were run using WINSTEPS 3.75.1 (Linacre, 2011) for the total scale and for each factor of engagement found in the EFA. As unidimensionality is an assumption of Rasch analyses, a principal component analysis (PCA) of the residuals of the total scale and of each engagement subscale were conducted. To infer unidimensionality, the eigenvalue of the first contrast, which explains the largest amount of variance in the residuals, should be less than two (Linacre, 2011). Additionally, item and person separation reliability, fit statistics, and the average measure of ability at each category, were examined for each engagement subscale. Item separation reliability refers to how well the test distinguishes between items along the measured variable, while person separation reliability refers to how well persons can be differentiated on the measured variable (Bond and Fox, 2007). These reliability coefficients can be interpreted in the same way as Cronbach's alpha.

Fit statistics help identify how well the data fit the model. Misfitting items or persons can illuminate departures from unidimensionality requirements or from the predicted model (Smith, Conrad, Chang, and Piazza, 2002). Mean fit squares (MNSQ) were examined to determine how well the data fit the model. MNSQs range from 0 to infinity and have an expected value of 1.0. The MNSQ is not symmetrically distributed about 1.0, as extreme values occur less frequently below 1.0 than above it (Smith, Schumacker, and Bush, 1998). Therefore, we calculated the percent of items in each subscale that were above critical values for both weighted (infit) and unweighted (outfit) mean squares and *z*-statistics. Following the recommendations of Smith et al. (1998), the critical values that were used to calculate the percent of items with extreme values were MNSQs greater than 1.1, 1.2, and 1.3 or less than .9, .8, and .7, and *z*-statistics greater than 2, 3, and 4, or less than −2, −3, and −4. Items that fell into the range of extreme values were considered for deletion.

Person-measure by category and category frequencies were examined to assess the effectiveness of the 4-point scale. The average person measure at each category should increase as the response category increases, indicating that as a person's latent trait (e.g., engagement) increases, there is a greater likelihood of them selecting a higher response category (Linacre, 1995). Category frequencies demonstrate the distribution of responses across each category. If a subscale had a category with few responses, response categories were collapsed, and the model was re-run with fewer response categories.

In order to analyze how well each subscale worked at covering various levels of engagement in math, person-item maps and item measures were examined. Person-item maps show the distribution of person and item measures along the logit scale and indicate the extent to which the items target the various levels of the person measures. Item measures allow for comparisons between items, by illustrating which items within a subscale were easiest, and most difficult, for a person to endorse.

Results

Exploratory Factor Analysis

All variables were transformed by taking the square-root of the inverse of the variable prior to analysis, due to moderate negative skewness for several variables. These transformations resulted in univariate normality for each variable, as evaluated through visual inspection of histograms, and the examination of skewness and kurtosis, which were well within acceptable limits (<|1|). After the transformations, there were no univariate outliers. Mahalanobis distance revealed three multivariate outliers ($\chi^2(18) > 42.31$, $p < .001$). These were deleted one at a time, leaving a sample size of $n = 357$.

Principal axis factor (PAF) extraction was performed on the 18 student-reported SEMS variables. Cattell's (1966) scree plot and Horn's (1965) parallel analysis support a three-factor model. Therefore, PAF extraction was performed specifying a three-factor solution. The three-factor model accounted for 48.53% of the total observed score variance. Varimax (orthogonal) rotations were examined as some variable loaded onto multiple factors. Variables with loadings less than .35 on all three factors were deleted (items 2, 3, 9, and 12). These items seem to represent some latent trait different from engagement. For instance, item 12, which states "I didn't answer my teacher's questions in math today, because I thought I might be wrong" seems likely to be measuring anxiety or self-confidence instead of engagement in math. The resulting rotated structure matrix revealed a pattern consistent with simple structure, with all retained variables having appreciable factor loadings of at least .42 (Table 1). The 14 items that were retained for analysis loaded onto the three separate dimensions of engagement: emotional, social, and cognitive. Correlations between items ranged from .44 to .68 for the emotional engagement subscale, .35 to .47 for the social engagement subscale, and .19 to .53 for the cognitive engagement subscale. The three-factor model accounts for 57% of the total observed score variance.

Rasch Analysis

Rasch analysis, in the form of a PCM, was conducted for the total 14-item engagement scale. The PCA of the residuals showed that the scale was multidimensional, as the first contrast had an eigenvalue of 2.8 (12% of the unexplained variance). As unidimensionality is an assumption of the PCM, individual PCMs were conducted separately for the dimensions of emotional engagement, social engagement, and cognitive engagement found in the EFA. The results of these analyses are divided into three subsections, each addressing one of the purposes of running the Rasch analysis. The first section (item fit) addresses how well the data conform to the PCM for each subscale. The second subsection (subscale thresholds) assesses the quality of the 4-point scale. The final subsection (item to person targeting) analyzes how well the scales work in covering different levels of engagement.

Item fit. Within each of the three subscales of engagement, unidimensionality, fit statistics, item and person separation reliability, and increasing measures of ability across categories were examined in order to assess how well the data fit the model. Table 2 provides an overview of fit by giving the descriptive statistics of the weighted (infit) and unweighted (outfit) mean square fit indices for each engagement subscale. The fit results for each individual subscale are described below.

Emotional Engagement. This five-item subscale can be considered unidimensional as the Rasch dimension accounted for 55.4% of the raw variance (36.1% persons and 19.3% items). The first contrast had an eigenvalue of 1.4 (12.6% of the unexplained variance). Person separation reliability for this five-item subscale was .74 for the 259 non-extreme persons, and .66 for all 357 respondents. Item separation reliability was .91.

Two of the items (11 and 17) were above the weighted MNSQ critical value of 1.1 (Table 3). Item 17 also had a weighted *z*-score above the critical value of 2 (Table 4). According to the guidelines suggested by Smith et al. (1998), with a sample size of 357 participants, 1.11 is the guideline for which a weighted MNSQ should be considered misfitting, while 1.32 is the guideline for which an unweighted MNSQ should be considered misfitting. However, as the MNSQs for these two items were close to the suggested cutoff values, and for all items, average ability increased across all categories, we decided to retain these items for the current analysis as there are currently no definitive answers for deciding on model fit (Embretson and Reise, 2000). However, caution for future studies should be noted.

Social Engagement. All items in this subscale had good fit (Table 3). Average ability increased across categories for every item. The Rasch dimension accounted for 49.3% of the raw variance (24.6% persons and 24.7% items). The first contrast had an eigenvalue of 1.4 (18.3% of the unexplained variance). No items were deleted from this subscale. Person separation reliability for this four-item subscale was .62 for the 320 non-extreme persons, and .66 for all 357 respondents. Item separation reliability was .98.

Cognitive Engagement. The original person separation reliability of this subscale was .43 for the 268 non-extreme persons, and .38 for all 357 respondents. Item separation reliability was .96. As part of the investigation of the reasons for this low person separation reliability, we first looked at the item fit. Item 4 "Today I only paid attention in math when it was interesting" had a weighted MNSQ of 1.33, and a *z*-score of 3.3. Upon further examination, we discovered that not all the response categories had a probability of being selected. Collapsing categories together did not help with item fit. Therefore, we decided to remove item 4 from the measure. The reason for the misfit of this item is likely due to the wording of the question, which is asking both about paying attention and about math being interesting. After the removal of item 4, the data were recalibrated. The modified 4-item subscale

Table 1

Factor Loadings for Exploratory Factor Analysis With Varimax Rotation of the Student Engagement in Mathematics Scale (SEMS)

	Items	Emotional Engagement	Social Engagement	Cognitive Engagement
10.	Math class was fun today	**.83**	.13	.12
11.	Today I felt bored in math class[1]	**.68**	−.04	.25
13.	I enjoyed thinking about math today	**.72**	.25	.23
16.	Learning math was interesting to me today	**.70**	.12	.32
17.	I liked the feeling of solving problems in math today	**.55**	.17	.27
5.	Today I talked about math to other kids in class	.10	**.58**	.15
6.	Today I helped other kids with math when they didn't know what to do	.12	**.68**	.06
7.	Today I shared ideas and materials with other kids in math class	.08	**.60**	.15
8.	Students in my math class helped each other learn today	.08	**.61**	.05
1.	Today in math class I worked as hard as I could	.34	.20	**.42**
4.	Today I only paid attention in math when it was interesting[1]	.07	−.016	**.42**
14.	Today it was important to me that I understood the math really well	.33	.19	**.53**
15.	I tried to learn as much as I could in math class today	.28	.25	**.66**
18.	I did a lot of thinking in math class today	.31	.19	**.48**

[1] Items 4 and 11 were reverse scored.

Note. $n = 357$. Factor loadings greater than .40 are in boldface. See Appendix for full SEMS questionnaire.

Table 2

Weighted and Unweighted MNSQ Descriptive Statistics

	Mean	*SD*	Min	Max
Emotional Engagement				
Infit	1.01	.18	.78	1.25
Outfit	.97	.18	1.26	.79
Social Engagement				
Infit	1.01	.05	.92	1.07
Outfit	.98	.04	.91	1.03
Cognitive Engagement (original)				
Infit	1.01	.18	.77	1.33
Outfit	.98	.14	.74	1.18
Cognitive Engagement (modified)				
Infit	.99	.09	.83	1.06
Outfit	.98	.11	.80	1.05

Note. *SD* is standard deviation. Infit refers to weighted MNSQs, while outfit refers to unweighted MNSQs. The original Cognitive Engagement subscale has five items, while the modified Cognitive Engagement subscale has four items.

Table 3

Mean Square (MNSQ) Frequency of Extreme Values

	Emotional Engagement		Social Engagement		Cognitive Engagement (original)		Cognitive Engagement (modified)	
	I	O	I	O	I	O	I	O
% > 1.3	0	0	0	0	20	0	0	0
% > 1.2	20	20	0	0	20	0	0	0
% > 1.1	40	20	0	0	20	20	0	0
% < 0.9	40	60	0	0	20	20	25	25
% < 0.8	20	20	0	0	20	20	0	0
% < 0.7	0	0	0	0	0	0	0	0

Note. **I** stands for *infit* (weighted) MNSQs, **O** stands for *outfit* (unweighted) MNSQs. Percentages are high because there are very few items in each subscale. The emotional engagement subscale has 5 items, the social engagement subscale has 4 items, the original cognitive engagement subscale had 5 items and the modified cognitive engagement subscale has 4 items.

had MNSQ values within the range recommended by Smith et al (1998).

In the four-item modified subscale, the Rasch dimension accounted for 44.4% of the raw variance (28.4% persons and 16.0% items). The first contrast had an eigenvalue of 1.5 (21.5% of the unexplained variance). Item separation reliability was .89, which is a slight decrease from the initial model. However, person separation reliability was .52, which is better fitting than the initial model, helping justify the removal of item 4. Though the person separation reliability was slightly improved over the initial model, it is still low. One way to improve this is by creating additional, effective items. Future work on this scale should involve the creation and testing of more items that seek to measure cognitive engagement.

Category frequencies. Table 5 presents the category frequency counts for each engagement subscale. In the Emotional and Social Engagement subscales, all four category responses are well-utilized. In the modified Cognitive Engagement subscale many more respondents selected responses that represented higher levels of cognitive engagement.

Item to person targeting. In order to examine how good the SEMS is at targeting persons with different levels of ability, person-item maps were assessed for each subsection of engagement. Additionally, item measures were examined (Table 6).

Figure 1 shows the person-item map for the Emotional Engagement subscale. The items cover a large part of the range of emotional engagement. However, there are a lack of items detecting high levels of emotional engagement. Item measure locations range from −0.46 to 0.61 (Table 6). Item 11 "Today I felt bored in math class" was the easiest for students to endorse, while item 13 "I enjoyed thinking about math today" was the most difficult.

Table 4

z-statistic (ZSTD) Frequency of Extreme Values

	Emotional Engagement		Social Engagement		Cognitive Engagement (original)		Cognitive Engagement (modified)	
	I	O	I	O	I	O	I	O
% > 4.0	0	0	0	0	0	0	0	0
% > 3.0	0	0	0	0	20	0	0	0
% > 2.0	20	20	0	0	20	0	0	0
% < −2.0	20	20	0	0	20	20	0	0
% < −3.0	0	0	0	0	0	0	0	0
% < −4.0	0	0	0	0	0	0	0	0

Note. **I** stands for *infit* (weighted) MNSQs, **O** stands for *outfit* (unweighted) MNSQs.
Percentages are high because there are very few items in each subscale. The emotional engagement subscale has 5 items, the social engagement subscale has 4 items, the original cognitive engagement subscale had 5 items and the modified cognitive engagement subscale has 4 items.

Table 5

Category Frequency Counts for each Engagement Subscale

	Observed Count		
Category	Emotional Engagement	Social Engagement	Cognitive Engagement
1	71	297	20
2	255	293	108
3	441	345	416
4	982	472	861

Table 6

Item Measures for Each Engagement Subscale

Item	Measure
Emotional Engagement	
10. Math class was fun today	−.17
11. Today I felt bored in math class[1]	−.46
13. I enjoyed thinking about math today	.61
16. Learning math was interesting to me today	.05
17. I liked the feeling of solving problems in math today	−.02
Social Engagement	
5. Today I talked about math to other kids in class	.07
6. Today I helped other kids with math when they didn't know what to do	.46
7. Today I shared ideas and materials with other kids in math class	.25
8. Students in my math class helped each other learn today	−.78
Cognitive Engagement	
1. Today in math class I worked as hard as I could	−.30
14. Today it was important to me that I understood the math really well	.02
15. I tried to learn as much as I could in math class today	−.28
18. I did a lot of thinking in math class today	.56

Figure 2 shows the person-item map for the Social Engagement subscale. The four items did well at discerning persons at the low to mid-levels of social engagement. However, there is a lack of items detecting very low or very high levels of social engagement. Item locations range from −0.78 to 0.46 (Table 6). Item 8 "Students in my math class helped each other learn today" was the easiest for students to endorse, while item 6 "Today I helped other kids with math when they didn't know what to do" was the most difficult.

Figure 3 shows the person-item map for the modified Cognitive Engagement subscale. The four items cover a wide range of cognitive engagement. However, more items are needed to differentiate between students with higher levels of cognitive engagement. Item locations range from −0.30 to 0.56 (Table 6). Item 1 "I tried to learn as much as I could in math class today" was the easiest for students to endorse, while item 18 "Today in math class I worked as hard as I could" was the most difficult.

Conclusions and Discussion

Many researchers have demonstrated positive associations between student engagement in school and academic outcomes (Christenson et al., 2012; Fredricks et al., 2004). Having a measure that specifically assesses engagement in mathematics is valuable as mathematics achievement represents an area of concern in the United States. The SEMS is such a measure that, to this point, has only been evaluated using *true score equation* techniques. This is the first study to investigate the quality of the SEMS using Rasch analysis.

During the exploratory factor analysis of the SEMS on the current data set, four items were deleted. The remaining items loaded onto one of the factors of Emotional Engagement, Social Engagement, or Cognitive Engagement. During the Rasch analysis, no items were deleted from the Emotional Engagement or Social Engagement subscales. One item was deleted from the Cognitive Engagement subscale. This modified subscale can be considered a more valid measure of the underlying cognitive engagement construct. Therefore, the results suggest that instead of administering the original 18-item SEMS to students, a reduced 13-item instrument would be more useful (see Appendix.) This shortened instrument might also be better suited for fifth graders.

The quality and utility of the 4-point ordered response scale of the SEMS was assessed through the PCM. Participants used all four response

categories equally for the Social Engagement subscale. In the other two engagement subscales the response category *No, not at all true* was endorsed by many fewer participants than the other three response categories. Future administration of the SEMS should be used to examine whether a 3-point response scale to the Emotional Engagement and Cognitive Engagement items has better fit.

The items on each engagement subscale seem to address respondents with a range of ability levels, with the main exception being a general absence of items in the higher ability range. This may not be the best measure for discerning different levels of engagement between students who have very high levels of engagement in mathematics.

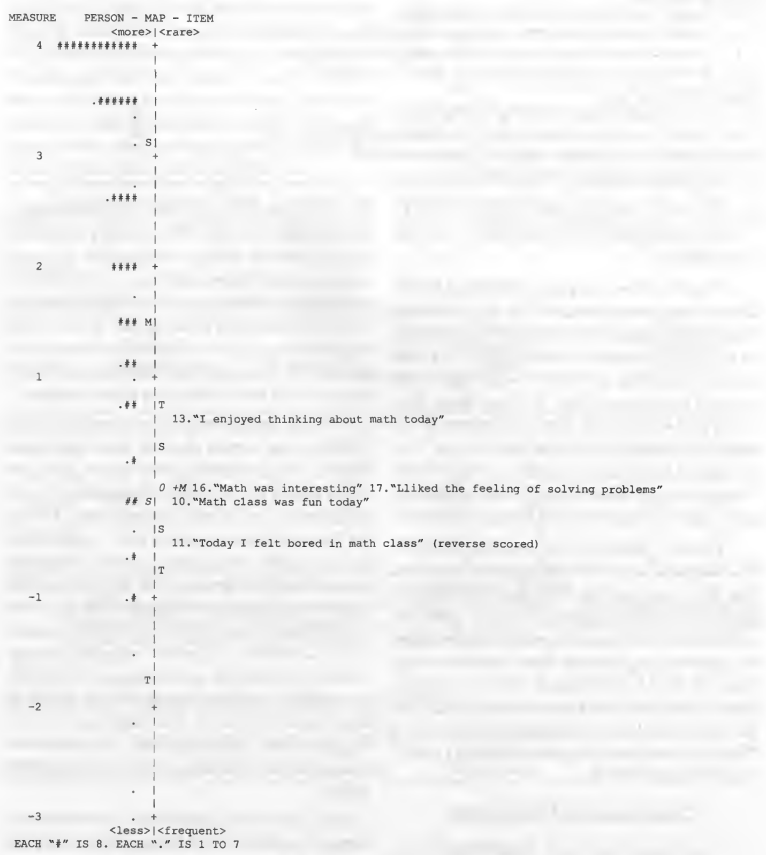

Figure 1. Person-item map for Emotional Engagement subscale

Future research on the SEMS should include a differential item functioning (DIF) analysis with a larger sample size of participants in order to see if one group is endorsing an item more easily than another group after controlling for ability (i.e., the latent trait of engagement) (Bond and Fox, 2007; Fischer and Molenaar, 1995). DIF analysis should especially be conducted by gender as this variable has been linked to engagement with boys showing lower levels of behavioral and emotional engagement than girls in the elementary and middle school years (Kindermann, 2007; Marks, 2000; Ponitz, Rimm-Kaufman, Brock, and Nathanson, 2009). Additionally, further research on the SEMS

Figure 2. Person-item map for Social Engagement subscale

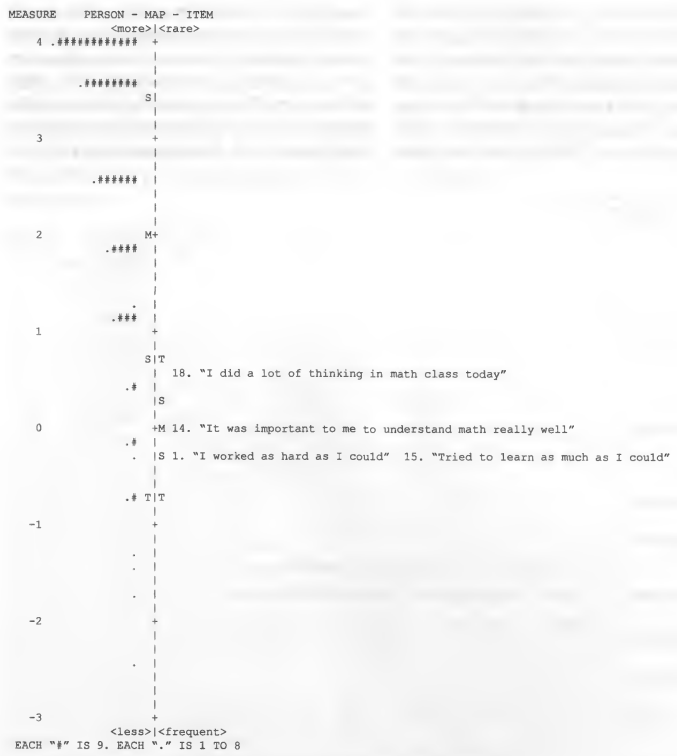

Figure 3. Person-item map for modified Cognitive Engagement subscale

using a large sample of students from different grade levels is needed before conclusions can be made about the generalizability of the ability of the SEMS to measure math engagement in other grade levels. However, with the deletion of several items (2, 3, 4, 9, and 12), the SEMS stands as a valid measure for analyzing the self-reported engagement in mathematics for fifth graders.

Acknowledgement

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B090002 and by the National Science Foundation under Grant No. DRL-0814872 to the University of Virginia. The opinions expressed are those of the authors and do not necessarily represent views of the Institute of Education Science, the U.S. Department of Education, or the National Science Foundation.

References

Appleton, J. J., Christenson, S. L., Kim, D., and Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, *44*, 427-445.

Bond, T. G., and Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Carini, R. M., Kuh, G. D., and Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, *47*, 1-32.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Christenson, S., Reschly, A. L., and Wylie, C. (2012). *Handbook of research on student engagement*. New York, NY: Springer.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Finn, J. D., and Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, and C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 97-131). New York, NY: Springer.

Fischer, G. H., and Molenaar, I. W. (Eds.) (1995). *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer.

Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*, 59-109.

Green, K. E., and Frantom, C. G. (2002, November). *Survey development and validation with the Rasch model*. Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC.

Greenwood, C. R., Horton, B. T., and Utley, C. A. (2002). Academic engagement: Current perspectives on research and practice. *School Psychology Review*, *31*, 328-349.

Griggs, M. S., Rimm-Kaufman, S. E., Merritt, E. G., and Patton, C. L. (2013). The Responsive Classroom approach and fifth grade students' math and science anxiety and self-efficacy. *School Psychology Quarterly*, *28*, 360-373.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Hughes, J., and Kwok, O. (2007). Influence of student-teacher and parent-teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *Journal of Educational Psychology*, *99*, 39-51.

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0 [Computer software]. Armonk, NY: IBM Corp.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187-200.

Kindermann, T. A. (2007). Effects of naturally existing peer groups on changes in academic engagement in a cohort of sixth graders. *Child Development*, *78*, 1186-1203.

King, J., and Bond, T. (1996). A Rasch analysis of a measure of computer anxiety. *Journal of Educational Computing Research*, *14*, 49-65.

Kong, Q. P., Wong, N. Y., and Lam, C. C. (2003). Student engagement in mathematics: Development of instrument and validation of construct. *Mathematics Education Research Journal*, *15*, 4-21.

Ladd, G. W., and Dinella, L. M. (2009). Continuity and change in early school engagement: Predictive of children's achievement trajectories from first to eighth grade? *Journal of Educational Psychology*, *101*, 190-206.

Linacre, J. M. (1995). Categorical misfit statistics. *Rasch Measurement Transactions*, *9*, 450-451.

Linacre, J. M. (2011). Winsteps® (Version 3.70.0) [Computer software]. Beaverton, OR: Winsteps.com.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690-700.

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, *37*, 153-184.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100-117.

McMahon, S. D., Parnes, A. L., Keys, C. B., and Viola, J. J. (2008). School belonging among low-income urban youth with disabilities: Testing a theoretical model. *Psychology in the Schools*, *45*, 387-401.

Meece, J. (2009). *Measure of student cognitive engagement*. Unpublished measure, University of North Carolina, Chapel Hill, NC.

Northeast Foundation for Children (2014). *Principles and practices of responsive classroom*. Retrieved May 3, 2014 from: http://www.responsiveclassroom.org/principles-and-practices-responsive-classroom

Patrick, H., Ryan, A. M., and Kaplan, A. (2007). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology*, *99*, 83-98.

Ponitz, C. C., Rimm-Kaufman, S. E., Brock, L. L., and Nathanson, L. (2009). Early adjustment, gender differences, and classroom organizational climate in first grade. *The Elementary School Journal*, *110*, 142-162.

Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., and Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review*, *38*, 102-120.

Program for International Student Assessment (2013). *PISA 2012 results*. Retrieved from: http://www.oecd.org/pisa/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Raykov, T., and Marcoulides, G. A. (2011). *Introduction to psychometric theory*, New York, NY: Springer.

Reeve, J., Jang, H., Carrell, D., Jeon, S., and Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, *28*, 147-169.

Reschly, A. L., and Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, and C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 3-19). New York, NY: Springer.

Rimm-Kaufman, S. E., Baroody, A. E., Larsen, R. A., Curby, T. W., and Abry, T. (2015). To what extent do teacher-student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning? *Journal of Educational Psychology*, *107*, 170-185.

Rimm-Kaufman, S. E., Larsen, R. A., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., et al. (2014). Efficacy of the Responsive Classroom approach: Results from a 3-year, longitudinal randomized controlled trial. *American Educational Research Journal*, *51*, 567-603.

Rowley, S. J., Kurtz-Costes, B., Meyer, R., and Kizzie, K. (2009). *Engagement and self-concept during the transition to middle school: Gender and domain-specific differences in change in African American youth*. Unpublished manuscript, University of Michigan, Ann Arbor, MI.

Skinner, E. A., and Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, *85*, 571-581.

Smith, E. V., Jr., Conrad, K. M., Chang, K., and Piazza, J. (2002). An introduction to Rasch measurement for scale development and person assessment. *Journal of Nursing Measurement*, *10*, 189-206.

Teman, E. D. (2013). A Rasch analysis of the statistical anxiety rating scale. *Journal of Applied Measurement*, *14*, 414-434.

Appendix

Student Engagement in Mathematics Scale (SEMS)

1. Today in math class I worked as hard as I could.
2. I thought about other things instead of math in math class today.[1]
3. I went back over things I didn't understand today in math class.
4. Today I only paid attention in math when it was interesting.[1]
5. Today I talked about math to other kids in class.
6. Today I helped other kids with math when they didn't know what to do.
7. Today I shared ideas and materials with other kids in math class.
8. Students in my math class helped each other learn today.
9. I raised my hand to answer questions in math class today.
10. Math class was fun today.
11. Today I felt bored in math class.[1]
12. I didn't answer my teacher's questions in math today, because I thought I might be wrong.[1]
13. I enjoyed thinking about math today.
14. Today it was important to me that I understood the math really well.
15. I tried to learn as much as I could in math class today.
16. Learning math was interesting to me today.
17. I liked the feeling of solving problems in math today.
18. I did a lot of thinking in math class today.

*All items scored on a four-point scale: 1 = *No, not at all true*; 2 = *A little true*; 3 = *Often true*; 4 = *Yes, very true*.

[1]Reverse-scored items

Note. Items #5, 6, 7, and 8 were adapted from Patrick, Ryan, and Kaplan's (2007) social engagement measure. Other items were developed based upon measures created by Meece (2009); Kong et al. (2003); Rowley, Kurtz-Costes, Meyer, and Kizzie (2009); and Skinner and Belmont (1993). Results from the PCM suggest deleting items 2, 3, 4, 9, and 12 in future administration of the questionnaire.